Coarse Vector Response Regression for long-term monitoring of ecosystems: revealing the causes of regime shifts in a brackish lagoon.

Claude Manté[a,*], Guillaume Bernard[b], Jean-Pierre Durbec[a]

[a] *Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, Campus de Luminy, Case 901, F13288 Marseille Cedex 09, France. tel: (+33) 486 090 631 fax: (+33) 486 090 641*

[b] *GIPREB Syndicat Mixte, Cours Mirabeau, Berre l'Étang, France.*

Abstract

Cover is the most frequently used measure for vegetation surveys. Generally, it is coded with the Abundance/ Dominance Braun-Blanquet's integer code, giving rise to data which are generally considered as ordinal. Since each one of these integers is associated with a whole interval of values, we argue that these codes actually convey more information than that of a simple order, and develop this point of view, considering them as imprecise data.

To our knowledge, there is no ready-made method to investigate relationships between a vector of such responses and several explanatory variables. Consequently, we propose a three-step method for this purpose. These steps are: (1) randomly recover (through a probability associated with the assessor's subjectivity) possible "original numerical responses"; (2) cluster these numerical response vectors according to an appropriate metric, into an appropriate number of groups; (3) average every variable conditionally to the classifier associated with step two, giving rise to "per group regression functions". This method is applied to explain with hydrological variables the abundance variations of *Potamogeton pectinatus* in a brackish lagoon (the Berre lagoon, Provence, France). We reveal the main relationships between *P. pectinatus* cover and fresh water inputs, salinity and nutrient abundance (nitrate and phosphate); the obtained results are compared to those from Canonical Correspondence Analysis. The proposed method is also tested on artificial data, similar to the original cover data.

Keywords: Stochastic restoration, Nonparametric regression, Coarse data, Transferable Belief Model, Clustering, Braun-Blanquet score

*Corresponding author

Email addresses: claude.mante@mio.osupytheas.fr (Claude Manté), guillaume.bernard@gipreb.fr (Guillaume Bernard), claude.mante@mio.osupytheas.fr (Jean-Pierre Durbec)

1. Introduction

Cover is the most frequently used indicator for vegetation surveys, since it is not destructive and requires relatively little effort compared to other measurement of vegetation health (Chen et al., 2008a; Kim and Travers, 1997a). We will use here a " la Braun-Blanquet" cover score adapted for marine ecology (see Table 1), because the field of vision of divers is reduced (Marcos-Diego et al., 2000). This coding was adopted for studying the temporal variations of the cover of the Berre lagoon by *Potamogeton pectinatus* (from 1970 to 2004), in connection with annual measuring of ten hydrological variables. The *P. pectinatus* cover was annually sampled at 35 stations around the lagoon, while hydrological variables were only sampled at a single place, each year.

Explaining the variations of the cover of this lagoon by *P. pectinatus* from variations of hydrological variables is typically a regression problem, which is made special by the structure of the surveys, since

- the cover is coded on an ordinal scale (this important point will be discussed further)
- the coding process is affected by the subjectivity of the observer; in addition, it can be disrupted by environmental factors
- each survey is associated with a vector of 35 cover codes, instead of a single value of some target variable
- the spatial dependence between the observations cannot be taken into account (no repetition, no spatial structure for hydrological data).

While most authors consider cover data as ordinal and debate about the troubles this causes for data processing (Podani, 2005, 2007; Ricotta and Avena, 2006; Van der Maarel, 2007), we will adopt another position, considering that it has the richer structure of coarse data (Heitjan and Rubin, 1991). This will enable us to propose an original method (Coarse Vector Response Regression: CVRR) combining stochastic restoration with clustering and nonparametric regression, in order to investigate relationships between such a vector of cover codes and several explanatory variables. The performance of CVRR will be tested first on well-designed Monte Carlo simulations and then on the Berre Lagoon data.

2. Data description

In this section, we provide an outline of the available data: implementation of the *P. pectinatus* cover survey, list of the hydrological independent variables.

2.1. The Berre lagoon and P. pectinatus data compilation

The Berre lagoon (Provence, Southern France) is one of the largest Mediterranean coastal lagoons (155 km^2). In the late 19^{th} and early 20^{th} centuries, urban development and mainly petrochemical industrialization of the lagoon's

region resulted in a steady increase in pollution. Since 1966, the diversion of the Durance River towards the Saint-Chamas EDF hydroelectric power plant, and then into the lagoon resulted in (i) a heavy input of freshwater (up to seven times the volume of the lagoon per year), (ii) the decline of surface water salinity from $2.4-3.6\,\%$ to $0.1-2.2\,\%$, (iii) eutrophication and unstable ecological conditions (Nérini et al., 2000). In the years following the diversion of the Durance River into the lagoon, newly forced, isolated, patches of *P. pectinatus* have been reported (Riouall, 1972; Stora, 1976) while, in the same time, other species, as *Zostera* beds, were drastically reduced (Bernard et al., 2005, 2007). At the mid 1980s, *P. pectinatus* constituted extensive continuous belts, along the north-west shoreline of the lagoon, from a few centimeters below the mean water level to 1 m depth (Mossé and Mossé, 1985).

By means of a GIS database, we coupled historical data (1970 to 1996) and ground truth (for the recent years: 1998 to 2004) in an attempt to assess the patterns of change of the *P. pectinatus* cover over time and to connect them with changes in the lagoon environment, on the basis of quantitative data. The whole shoreline of the Berre lagoon was surveyed annually between 1998 and 2004 (from a small boat and by snorkeling) and *P. pectinatus* cover was recorded. Thirty five stations were identified and geographically localized as common sampling stations for each year of the whole time series (Figure 1). Historical data (for the years 1970, 1972, 1984, 1986, 1989) were then coupled with ground truth (for the years 1990, 1994, 1995, 1996, 1998, 2002, 2004) in a GIS database (ArcGIS 8.0®). For the whole time series, the observations were done during July and August. Of course, the cover is specific to the investigated

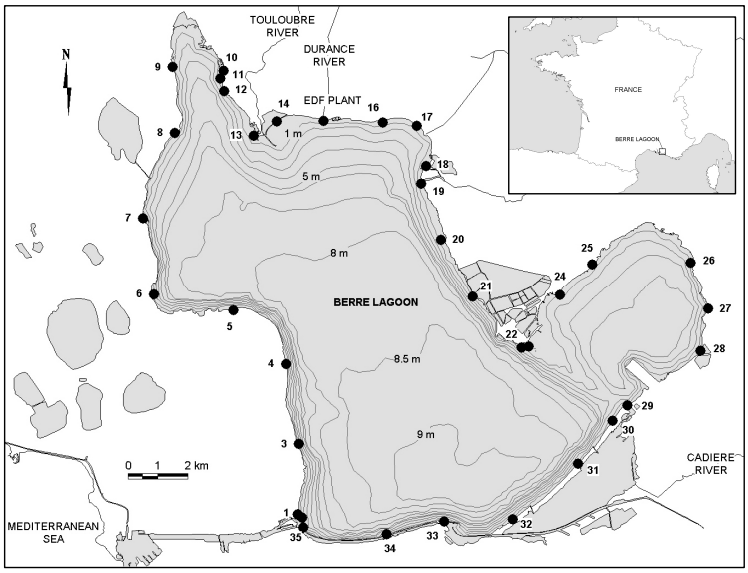


Figure 1:

sampling station. The surface area of each station has been defined as the surface area included between the 50 m long fraction of the shoreline centered on the sampling station. The records of the *P. pectinatus* cover have been coded from 0 to 4 (see Table 1).

Code	Observations	Corresponding interval for cover	Bound for cover (%)
0	absent	none	0
1	isolated shoots	negligible	$]0, 0.05[$
2	isolated patches	< 2	$[0.05, 2[$
3	confluent patches	$2 << 10$	$[2, 10[$
4	continuous belts	> 10	$[10, 35]$

Table 1: Codes and percentages corresponding of the cover of *P. pectinatus* in the sampling stations

2.2. A sketch of the temporal evolution of P. pectinatus - the regime shift hypothesis

In a preliminary step, the sum of the cover codes from all the 35 stations has been defined as a yearly total abundance index, denoted T. It can vary from 0 (*P. pectinatus* absent in the whole lagoon) to 140 (present with a cover $>$ 10 % in all the 35 stations). Between 1970 and 2004, T varied from a minimum of 0 to a maximum of 62, in 1986 (see Figure 2). This sudden year-to-year variations in *P. pectinatus* abundance suggests a "regime shift", defined as a rapid reorganization of the ecosystem from a stable state to another along a non linear evolution (Brock and Carpenter, 2006; Rodionov, 2004; Rodionov and Overland, 2005). In our case, the ecosystem switches between state (a): the species was present in very few places with very low abundance (sometimes limited to isolated shoots) and state (b): it constituted extensive beds along large parts of the lagoon shore. While the shift from (a) to (b) cannot be dated precisely due to lack of data, the shift back to (a) occurred between 1995 and 1996 (see Figure 2).

2.3. Hydrological data

The hydrological variables, and their codes for data analyzes, are mean annual salinity (coded **Sal.**); mean salinity in August (**Sal. Aug.**); total annual inputs of freshwater by the Durance river from January to December (Mt/a)(**FW**); total inputs of freshwater by Durance River from October to September (Mt/a) (**FW-1**); annual inputs of silts from the Durance River (10^3t/a) (**Silt**); annual inputs of P-PO_4 (t) (**P- PO_4**); annual inputs of N-NO_3 (t) (**N-NO_3**); mean concentration of N-NO_3 in surface water (μmol.L^{-1}) ([**N-NO_3**]); mean concentration of P-PO_4 in surface water (μmol.L^{-1}) ([**P-PO_4**]); concentration of suspended solids in surface water (wg.L^{-1}) ([**SS**]).

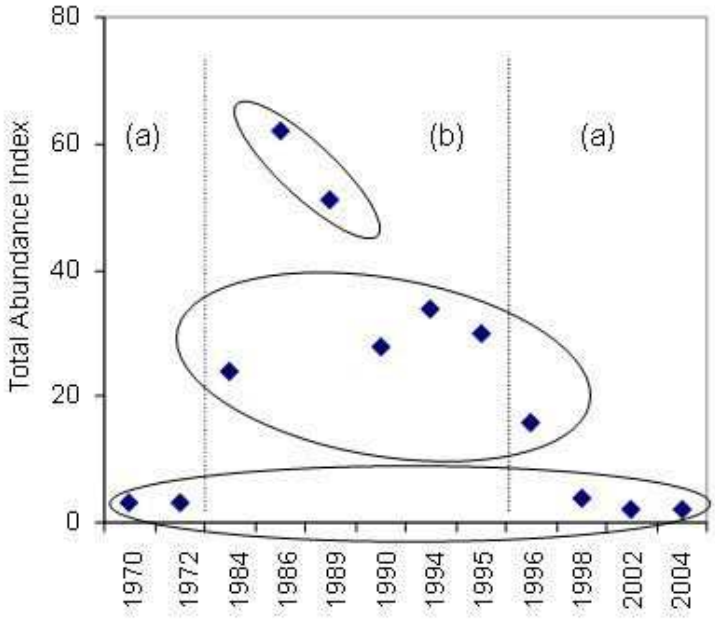


Figure 2:

3. Statistical methods

In this section, we will show the shortcomings of two classical methods for processing our data, and propose an original regression method for coarse vector responses designed for our purpose.

3.1. Standard analyzes

3.1.1. Basic linear regression

The exploration of relationships between *P. pectinatus* cover (using total abundance index T) and hydrological conditions of the lagoon was first carried out through simple linear models. These basic analyzes had disastrous outcomes, exemplified on Figure 3 by the "best" result: concentrations in $N\text{-}NO_3$ ($R^2 = 0.186$, P-value of the regression slope = 0.16).

But notice T is a very rough summary of these 35 values which are not even authentic numbers, only ordinal codes. Furthermore, one can infer from Figure 3 that the relationships we are trying to reveal are likely nonlinear.

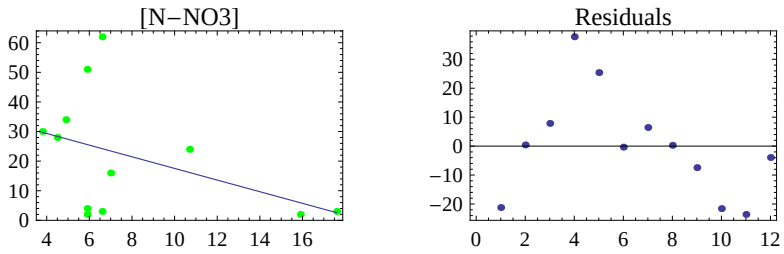


Figure 3:

3.1.2. Canonical Correspondence Analysis (CCA)

To investigate relationships between *P. pectinatus* cover and hydrological variables, we then considered a more efficient method: CCA (Greenacre, 2007; Ter Braak, 2006; Ter Braak and Venderschot, 1995). It is a multivariate exploratory method designed for investigating the relationships between rows (individuals) and columns (characters) of a contingency table, in connection with the structure of a second table of numerical variables measured on the same individuals. It is in a way a generalization of simple correspondence analysis where linear restrictions are set on attributed row and column scores. Carrying on CCA needed to associate the five-component vector of counts of the ordinal codes of *P. Pectinatus* cover observed at the 35 sites to each observation (year). These counts were displayed on a 12 by 5 frequency table form. This table was associated to the 12 by 10 table of values of the hydrological variables. In the present analysis the row scores are linear combination of the hydrological variables (mufti-linear regression); the columns scores being computed by weighting averaging. Thus, the variations of *P. pectinatus* cover could be directly related to the variations of hydrological variables.The results can be displayed on ordination plots simultaneously visualizing the correlations between hydrological variables and the similarities between observations (years) and degrees of cover. The `cca` procedure from the `R package ‘‘anacor’’` was chosen for calculations (De Leuw and Mair, 2009).

The two first axes (Figure 4) corresponded to about 80% of total chi-square. The first dimension (64%) was clearly more important than the second dimension (18%) for explaining the variability of data. Only four hydrological variables were highly or moderately correlated with the two first axes of cca. These variables and their correlation coefficients with the two axes were: [N-NO_3] (r_1=-0.446, r_2=-0.110), Sal.Aug. (r_1=-0.171, r_2=0.694), FW-1 (r_1=0.237, r_2=-0.415), and [P-PO_4] (r_1=-0.03, r_2=0.685).

The observations (years) could be clustered in three groups (see Figure 4). The first group was not very dispersed; it was characterized by a “0” degree of cover and included 1970, 1972, 1996, 1998, 2002, and 2004. It was essentially associated with increased values of [N-NO_3], and at a lesser degree to Silt and Sal.Aug (r=-0.17 with first axis). The second group included 1984, 1990, 1994,

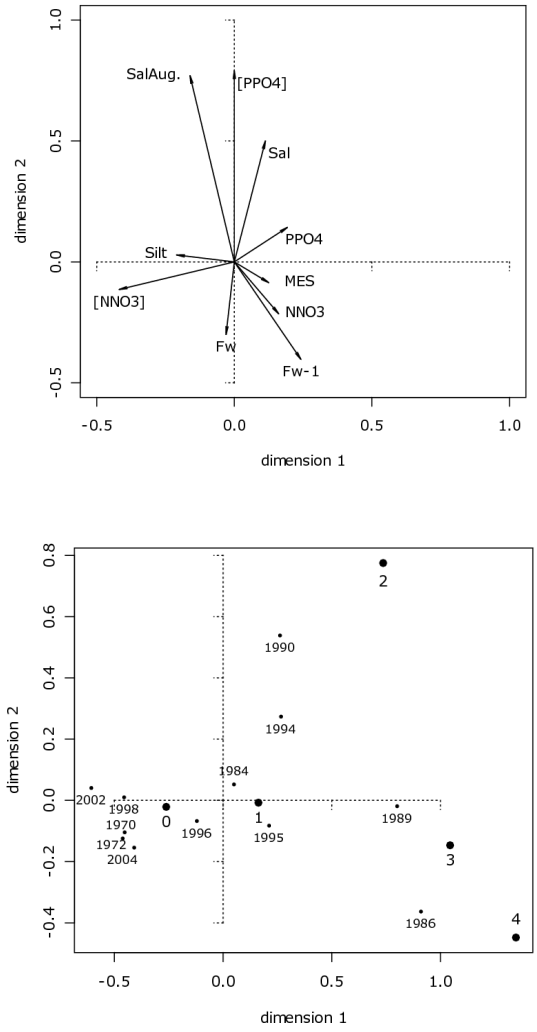


Figure 4:

1995, and was rather characterized by the degree of cover “1”. The position of this group on the first dimension, showed that it was rather associated with low

values of [N-NO_3] (r_1=-0.446) and had a weak association with increased values of FW-1 (r_1=0.237).This group was very scattered along the second dimension. It was an heterogeneous group. At last, the third group included 1986 and 1989 and was characterized by the degrees "3" and "4" of cover, increased values of FW-1 (r_1=0.237) and low values of Sal.Aug (r_2=0.694) and [P-PO_4] (r_2=0.685). The second dimension (axis 2) enabled us to characterize the degree of cover "2" by increased values of Sal. and [P-PO_4]. Notice that this classification is quite compatible with the regime shift hypothesis.

CCA was rather well-suited for our purpose, but it suffers from several drawbacks:

- the cover is considered as a categorical variable;
- it is a linear method, while the investigated relationships should be non-linear (see Figure 3 and associated comments);
- in the present study, where the number of hydrological variables (10) is close to the number of observations (12 years), CCA leads approximately to the same results as those of a correspondence analysis (CA) (Ter Braak and Venderschot, 1995).

Thus, both the above methods didn't take into account essential characteristics of the data. Classical methods of ordinal regression (Guizan and Harrell, 2000) cannot be used either, because in our case the dependent variable is a vector of ordinal codes instead of a single target variable. Consequently, we propose a specific method for analyzing such data. This method combines stochastic restoration, clustering and nonparametric regression, in a sequence of trials. A trial consists of three steps:

1. *(restoration)* replace each annual observation (a vector of 35 ordinal codes) by an appropriate vector of numbers in $[0, 1]$
2. *(classification)* split the obtained set of vectors into an appropriate **fixed number** G of classes
3. *(regression)* compute the cover mean versus the mean of each hydrological variable, **conditionally to the classifier**.

After performing a reasonable number R of trials (here: $R = 200$) one obtains, for each variable (including the cover), R different G-uples which give rise to "per group regression functions" (see Figures 8, 10, 13 &15). The method is exposed in more details hereunder.

3.2. The proposed method: CVRR

Consider the bulk "ecological state" of the lagoon, defined as the *P. pectinatus* extension, described for the k^{th} year by the random vector of ordinal codes $\overrightarrow{C^k} := \left\{{c_1}^k, \cdots, {c_{35}}^k\right\}$ - upper indices stand for years, while the lower ones stand for stations.

To highlight the hydrological variables responsible of the ecological state of the lagoon, we will firstly follow Singer et al. (2004). These authors proposed

to assign scores to ordinal categories before applying linear models and other methods designed for continuous variables, arguing that ordinal variables are indeed ill-suited for such models. In our case, since the ith modality, c_i, is associated with some interval of cover $[a_i, b_i[$ (see Table 1), a natural score is the "central score": $S(c_i) := \frac{b_i+a_i}{2} \in [0,1]$; Billard and Diday (2006), for instance, proposed this coding to define PCA for interval variables.

3.2.1. From ecological states to per group regression functions through the Central Score

Combining $\overrightarrow{C^k}$ with the central score S, we obtain the k^{th} "digitized state" $\overrightarrow{b}^k := \{S(c_1^k), \ldots, S(c_{35}^k)\}$ of the lagoon. We aim at explaining the variations of these states, keeping in mind that there are clusters of states (regime shift hypothesis) and that the influence of hydrological variables on the cover is likely non-linear.

Clustering the states with the Partitioning Around Medoids algorithm (PAM) (Kaufman and Rousseew, 1990), for instance, makes it possible to get rid of a major part of noise. This robust algorithm (Van der Laan et al., 2003; Kaufman and Rousseew, 1990) produces a set of cluster centers (medoids) from a dissimilarity matrix between objects of interest. When the "natural" number of clusters is unknown, the Silhouette test is one of the best methods for discovering it. Roughly speaking, it consists (Van der Laan et al., 2003) in optimizing the number of clusters in order that the clusters are homogeneous and stable (transferring an element from a cluster to another one would not improve the classification). We jointly used both these methods to cluster $\{\overrightarrow{b}^1, \ldots, \overrightarrow{b}^{12}\}$ into a convenient number of groups, from points of view associated with different distances between digitized states.

For quantifying the differences between successive ecological states, we had to choose an appropriate distance between these vectors of 35 attributed scores. Notice that the usual Euclidean distance implicitly takes into account the geographical position of the sampling sites, which corresponds to the 35 coordinates of each state vector $\overrightarrow{b^k}$. Consequently, if γ denotes a permutation of these coordinates, we have generally, for any pair (k, m) of digitized states: $\| \overrightarrow{b}^k - \overrightarrow{b}^m \| \neq \| \overrightarrow{b}^k - \gamma \circ \overrightarrow{b}^m \|$. Since only one value of each hydrological variable is associated with each state, independently of any geographical localization around the lagoon, such a property is undesirable. Consequently, distances between statistical distributions of annual scores were used. Amongst the numerous distances between distributions used by statisticians and probabilists (Gibbs and Su, 2002), we discarded those based on probability densities (Hellinger and Bhattacharya distances, divergences, *etc.*), because their computation needs a preliminary density estimation step. We focused instead on two distances based on distribution functions: the Kolmogorov distance:

$$DK(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|$$

and the Wasserstein (or Kantorovich) distance (Gibbs and Su, 2002):

$$DW(F_1, F_2) = \int_{-\infty}^{\infty} |F_1(x) - F_2(x)|dx.$$

Because both these distances gave very similar results, we will only focus on DW.

According to the Silhouette test, the set of states is optimally split into three groups of years. These groups are G_1:={1970,1972,1998,2002,2004},

G_2:={1984,1990,1994,1995,1996}, and G_3:={1986,1989}; they are quite compatible with the regime shift hypothesis (see Figure 2).

Consider now (formally) the nonparametric regression function associated with some variable Z and the digitized state: $m_Z(\overrightarrow{b}) := E(Z|\overrightarrow{B} = \overrightarrow{b})$. It is of course of no practical interest, but remember that $\{\overrightarrow{b}^1, ..., \overrightarrow{b}^{12}\}$ consist in three homogeneous groups. We can use the classifier $G\left(\overrightarrow{B}\right) \in \{1, 2, 3\}$ associating to each state its class number to build a valid regressor. It is the "per group regression function":

$$\widetilde{m}_Z(g) := E(Z|G(\overrightarrow{B}) = g).$$

In this setting, the conditional relationship between two variables X an Z can be summarized by the three points:

$$\{(\widetilde{m}_X(g_1), \widetilde{m}_Z(g_1)), (\widetilde{m}_X(g_2), \widetilde{m}_Z(g_2)), (\widetilde{m}_X(g_3), \widetilde{m}_Z(g_3))\}.$$

Fixing $Z := S(C)$, we get a schematic representation of the relationships between the digitized cover Z and each hydrological variable X, by displaying these expectations on a bi-dimensional scatterplot. Thus, while usual regression consists in finding a curve f such that $Z \approx f(X) + \varepsilon$, we merely plot the graph of the random variable (X, Z) coarsened (Heitjan and Rubin, 1991) by the classifier. Figure 5 illustrates the result obtained for X=[N-NO_3].

This plot clearly captures the information shown by Figure 3, but one could object that it depends on the score chosen. Shall we display a totally different graph by using another score? To control the dependence between the classifier and the score, we propose hereafter to supersede deterministic scores by well-suited random ones. This lead us to investigate the nature of cover further.

3.2.2. The cover: its evaluation and coding

Let $\pi \in [0, 1]$ be the actual (but unknown) *P. pectinatus* cover at one sampling site, for one sampling year. In practical terms, the diver first got an "estimation" $\hat{\pi}$ of this parameter by visual census (snorkeling); afterward, he coarsened this estimation by assigning it to some interval $O(\hat{\pi}) \in \{0, 1, 2, 3, 4\}$, giving rise to a record. Thus, even if the cover firstly appears as an ordinal

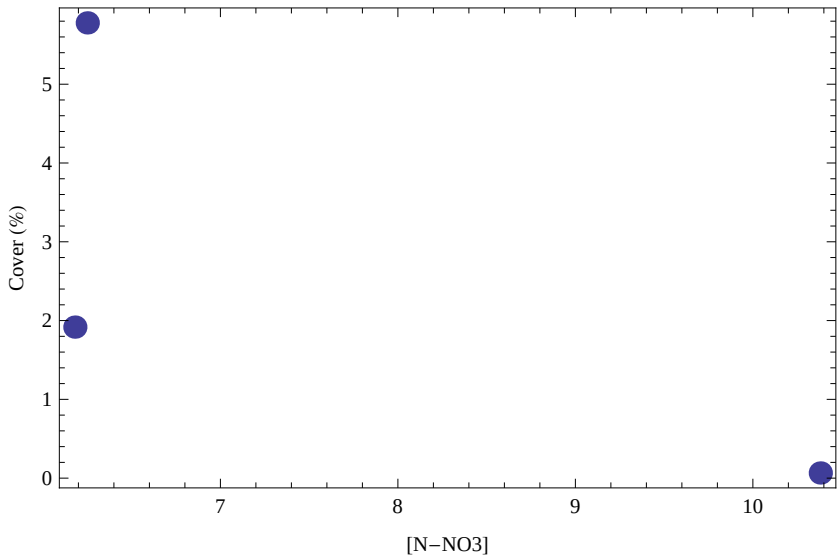


Figure 5:

variable, it conveys more information than just some order. A point of cardinal importance with π is indeed its imprecision: we cannot be sure that its actual value has been assigned to the right interval by the diver. This imprecision can be imputed to three different (but intricated) causes: noise, environmental factors, and subjectivity of the diver.

But notice that in our case, under no circumstances $\hat{\pi}$ results from a measurement! It is indeed the opinion of the diver about the cover of a parcel, a subjective belief integrating simultaneously environmental disruptions, which is afterward coarsened in accordance with the coding displayed in Table 1. Such data can be handled in the setting of the Evidence Theory (Shafer, 1976), widely used in the Artificial Intelligence community. In the next sections and in Appendices A&B, we propose a method for generating random covers, based on the Transferable Belief Model (TBM). This method makes it possible to involve subjective factors in the coarsening process and, consequently, in the restoration process.

In the nineties, Smets (Smets and Kennes, 1994; Smets, 1999) developed a model for representing quantified beliefs, the Transferable Belief Model, "supposed simulating the behavior of a reasonable and consistent agent" (Smets and Kennes, 1994). It is a subjectivist and non-probabilist extension of the Evidence Theory of Shafer (1976), where belief functions are interpreted as weighted opinions held by agents. The TBM is a two-levels model:

- the **credal** level, where subjective personal beliefs of the agents are entertained, upgraded, discounted, aggregated, *etc.* (Ha-Duong, 2008; Smets, 1990; Delmotte and Smets, 2004), see also (Shafer, 1976)

- the **pignistic** level (for betting), where beliefs are translated into probabilities to make decisions (Smets, 2005b).

Our case is indeed similar to the "one sensor problem" handled by Delmotte and Smets (2004), which consists in identifying a target from signals obtained from a sensor. These authors showed (see Appendix A) that this problem can be solved by computing pignistic probabilities depending on conditional probabilities obtained by training the sensor in a preliminary step. Since we don't possess such informations, we propose in Appendix B to work in a completely symbolic setting: we will use the TBM to build pignistic probability distributions associated with typical theoretical behaviors of the diver (**confident** or **suspicious**). These distributions will be used to restore the *P. pectinatus* cover observations with random scores.

3.2.3. Stochastic restoration of the cover

It is now time to describe the proposed stochastic restoration scheme. From the outset, we considered apart the first category, claiming that when *P. pectinatus* is not observed at all, $\pi = 0$.

Consider a diver $\mathcal{B}$, whose particular behavior has been formalized in the TBM setting (see Appendix B for detailed explanations); let $BetP_{\mathcal{B}}$ be the corresponding pignistic probability. Because of the coding used, this distribution is discrete but notice that π as well as its perception by $\mathcal{B}$, $\hat{\pi}$, are naturally continuous. Thus, $BetP_{\mathcal{B}}$ should be absolutely continuous (Smets (2005a) considered such pignistic probabilities). Consequently, as far as it is acceptable, we will supersede the original $BetP_{\mathcal{B}}$ by a convenient Beta distribution $\beta(p_{\mathcal{B}}, q_{\mathcal{B}})$. We chose this family of distributions because it proved its efficiency for modeling plant cover (Chen et al., 2008b,a; Irvine and Rodhouse, 2010). It is noteworthy that a Beta distribution is also a particular case of the BetaPERT distribution, frequently used to model expert opinion in the Risk Management literature (Paisley and Hostrup-Pedersen, 2004; Pellegrino and Costantino, 2012).

The vector of parameters $(p_{\mathcal{B}}, q_{\mathcal{B}})$ of the fitted distribution will result from minimizing $DW\left(FS_{BetP_{\mathcal{B}}}, F_{\beta(p,q)}\right)$, where $F_{\beta(p,q)}$ denotes the d.f. of some Beta distribution, and $FS_{BetP_{\mathcal{B}}}$ denotes the Stineman monotonic interpolation (Wagon, 2000) of the empirical d.f. associated with $BetP_{\mathcal{B}}$ (see Figures 7 , 9, 12 and 14).

Random scores will then be generated from the data by using the rule $\sigma_{\mathcal{B}} : C \longrightarrow [0,1]$ below:

$$\sigma_{\mathcal{B}} : \begin{cases} 0 \mapsto 0 \\ i \mapsto \beta(p_{\mathcal{B}}, q_{\mathcal{B}})|_{C_i} \quad if\ i \geq 1 \end{cases}$$

where $\beta(p_{\mathcal{B}}, q_{\mathcal{B}})|_{C_i}$ denotes the restriction of the probability to C_i. Notice that this restoration process depends on the conditions of observation (sensu lato), and not on the data, which is natural. By following the process described in Section 3.2.1, each set of restored states can be afterward clustered into an optimal number of groups determined by the Silhouette test. Finally, per group

regression functions are computed from the results of a reasonable number of trials.

4. Monte Carlo experiments

To put CVRR to the test, we performed simulations in conditions similar to those of the *Potamogeton* data. Suppose we can observe the values $\{\nu_1, \nu_2, ..., \nu_K\}$ of K "hydrological random variables" $\{V_1 \in \mathbb{V}_1, V_2 \in \mathbb{V}_2, ..., V_K \in \mathbb{V}_K\}$, and there are K known measurable functions, $\{\psi_1, \psi_2, ..., \psi_K\}$ determining the "cover":

$$\forall\ 1 \leq k \leq K, \quad \psi_k(v_k) = \pi \in [0,1].$$

We chose to define an ecological state as a random vector, whose independent coordinates obey a common law $\mathcal{D}_g$, for some $g \in \{1, \cdots, G\}$. Thus, we should find at most G clusters in a sample of states. But how could we simultaneously generate clusters of such vectors of cover (or "ecological states") and bulk values of the forcing hydrological variables, according to these relations?

4.1. Pitfalls of such simulations

A straightforward method for simulating a vector of cover would consist in generating a N-sample (here, N=35) of some probability distribution L_k associated with the kth variable, and consider the N-sample of covers induced on $[0,1]$ by ψ_k. This approach raises complicated issues, because the distributions $\{L_1, L_2, ..., L_K\}$ should be in coherence with the functions $\{\psi_1, ..., \psi_K\}$, *i.e.* the induced cover probabilities associated with different hydrological variables should be the same! In other words, for any pair (k, m) of indices, the induced laws $\psi_k * L_k$ and $\psi_m * L_m$ should obey a common distribution $\mathcal{D}$ on $[0,1]$. This is a severe constraint, made worse by another constraint: suppose we simulate this way T ecological states; the corresponding vectors should belong to at most G clusters!

A simpler way to proceed consists, when ψ_k possesses (in some sense) a measurable inverse $\varphi_k := \psi_k^{-1}$, in generating a sample $\{\pi_1, \cdots, \pi_{35}\}$ from some distribution $\mathcal{D}_g$ on $[0,1]$ associated with a cluster, and in computing $\{\varphi_k(\pi_1), \cdots, \varphi_k(\pi_{35})\}$. Notice that in this case $\{\varphi_k(\pi_1), \cdots, \varphi_k(\pi_{35})\}$ is a sample of the distribution $L_k^g := \varphi_k * \mathcal{D}_g$.

We will follow this strategy.

4.2. A straightforward case: simulating one-to-one relationships

Suppose ψ_k is bijective; we can merely use its inverse, $\varphi_k := \psi_k^{-1}$ for computing the random value of the kth hydrological variable for some vector of random covers, $\overrightarrow{\pi}$. Choosing the mean as a centrality criterion (it could be the median, as well), we can easily generate a realist bulk measure of V_k at the "time t": $\nu_k = \overline{\varphi_k}(\overrightarrow{\pi}^t) := \epsilon + \frac{1}{35}\sum_{i=1}^{35} \varphi_k(\pi_i^t)$, where $\epsilon \approx N(0, \sigma)$ is an additive noise. Thus, if all the possible relations were bijective, CVRR should enable us to recover any ψ_k from the codes and hydrological variables.

4.3. A general alternative: simulating a multifunction

But statistical relationships are not always one to one, and we must imagine more complicated situations (Einbeck and Tutz, 2006a,b) ! Suppose ψ_k is a multi-valued application (*i.e.* $\psi_k(\nu) \in \mathcal{P}([0,1])$ is not a singleton). Since ψ_k is not a function, it doesn't have an inverse, but we can nevertheless consider the associated inverse relation (Halmos, 1997), defined by $\Phi_k := \{(\pi, \nu) : \pi \in \psi_k(\nu)\}$. Alternatively, we can consider its lower inverse application (Berge, 1966), $\psi_k^-(\pi) := \{\nu : \nu \in \mathbb{V}_k, \pi \in \psi_k(\nu)\}$. Berge (1966) has shown that:

- if ψ_k is injective (*i.e.* $\nu^1 \neq \nu^2 \Longrightarrow \psi_k(\nu^1) \cap \psi_k(\nu^2) = \emptyset$), $\varphi_k := \psi_k^-$ is a function
- if ψ_k is a function, the application φ_k is injective
- $\varphi_k^- = \psi_k$.

In conclusion, as soon as ψ_k is injective, Φ_k is the graph of φ_k, and we can base our simulations on covers and φ_k.

The algorithm used for simulating a vector of covers and hydrological variables is thus:

1. generate a random vector $\overrightarrow{\pi} = (\pi_1, ..., \pi_{35})$, whose coordinates obey a given law $\mathcal{D}_g$, for some $g \in \{1, \cdots, G\}$
2. on the one hand, coarsen this vector according to some given table of codes, getting a row of a matrix M of ordinal codes
3. on the other hand, apply $\{\overline{\varphi}_1, ..., \overline{\varphi}_K\}$ to $\overrightarrow{\pi}$ and obtain a vector $(\nu_1, ..., \nu_K)$ of the so-called hydrological variables, getting a row of another matrix, W.

5. Results

5.1. Results of CVRR on simulated covers and $K = 6$ hydrological variables

We generated according to the algorithm above 30 random Gaussian vectors of cover, $\{(\pi_1^t, \cdots, \pi_{35}^t), 1 \leq t \leq 30\}$, such that all the components of the vector of covers at the "time" t obey a common distribution $N(\mu_t, \sigma)$. The mean μ_t was randomly drawn from the set $\{0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9, 1\}$, and σ was arbitrarily fixed to 0.1. Thus, each row of the obtained 30 by 35 matrix of covers, Π, belonged to some group of Gaussian vectors, and we had at the most $G = 8$ groups of covers.

5.1.1. Using the completed GIPREB grid

The simulated covers were first coarsened by using the intervals bounds of Table 1, completed by $]0.35, 1]$ (the cover is thus coded on 6 modalities), giving rise to a 30 by 35 matrix M of codes. Notice that this grid is very ill-suited for such data, as the reader can see on Figure 6.

The pignistic probability (see Appendix B) associated with the confident agent $\mathcal{B}_C$ is

$$BetP_{\mathcal{C}} = (0.1868967, 0.208193, 0.20982056, 0.208193, 0.1868967),$$

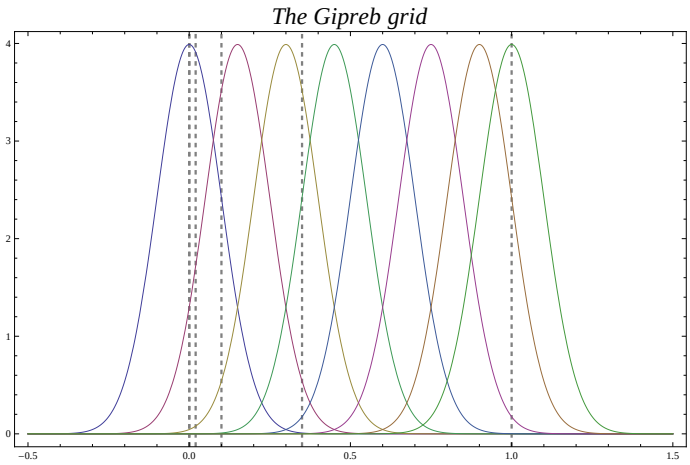


Figure 6:

and Figure 7 shows that the corresponding distribution function (d.f.) is well-fitted by the d.f. of $\beta\,(0.260144, 1.30305)$. Applying $\{\overline{\varphi}_1, ..., \overline{\varphi}_6\}$ to the rows of Π, we obtained a 30 by 6 matrix W. Then we processed this pair of matrices exactly the same way as with the Berre data in Section 3.2.1.

We plotted on Figure 8 the results obtained by $\mathcal{B}_C$ with two discontinuous applications (ψ_1 and ψ_2), two smooth bijective functions (ψ_4 and ψ_5) , and two injective (but multi-valued) applications, ψ_3 and ψ_6 (for thorough explanations about the graphical elaboration of this plot, see Section 5.2). Notice that none of the applications is correctly restored, excepted in the case of small cover values ($\pi < 0.5$).

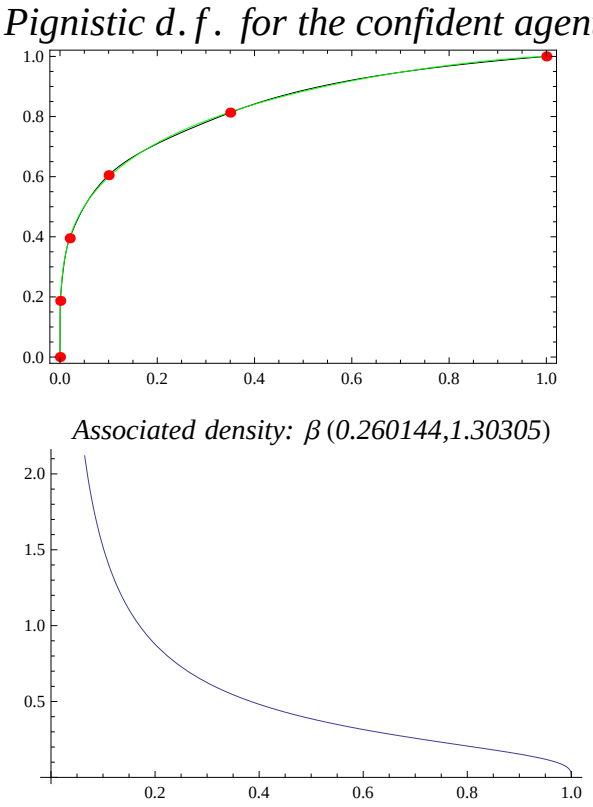


Figure 7:

Consider now the suspicious agent $\mathcal{B}_S$, whose pignistic probability is

$$BetP_{\mathcal{S}} = (0.0004964, 0.0207859, 0.0821761, 0.2504686, 0.6460731).$$

One can easily verify on Figure 9 that the corresponding d.f. is perfectly fitted by the distribution $\beta\,(0.9875, 0.9986)$, very close indeed to the uniform distribution.

This agent obtains results slightly better than those of $\mathcal{B}_C$ (see Figure 10); this is natural, since the GIPREB grid was not designed for such data! Nevertheless, the results of $\mathcal{B}_S$ are pretty feeble too, and this is also imputable to the grid. Suppose for instance we observed a very good ecological state: $\overrightarrow{Nice} = \{4_1, \cdots, 4_{35}\}$, and let $\{x_1, \cdots, x_{35}\}$ be associated simulated covers. Since the pignistic probability is approximately uniform (see Figure 9), the generated data are approximately uniform on $[0.35, 1]$. Since the mean and the standard deviation of the uniform distribution on $[0.35, 1]$ are respectively $\mu = 0.675$ and $\sigma = 0.18764$, the Gnedenko central limit theorem shows that $\bar{x}_{35} = \frac{\sum_{i=1}^{35} x_i}{35}$ approximately obeys $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{35}}\right) = \mathcal{N}\,(0.675, 0.03172)$. Thus, roughly speaking, we can say that $\bar{x}_{35}$ is very close to 0.675, and that it is

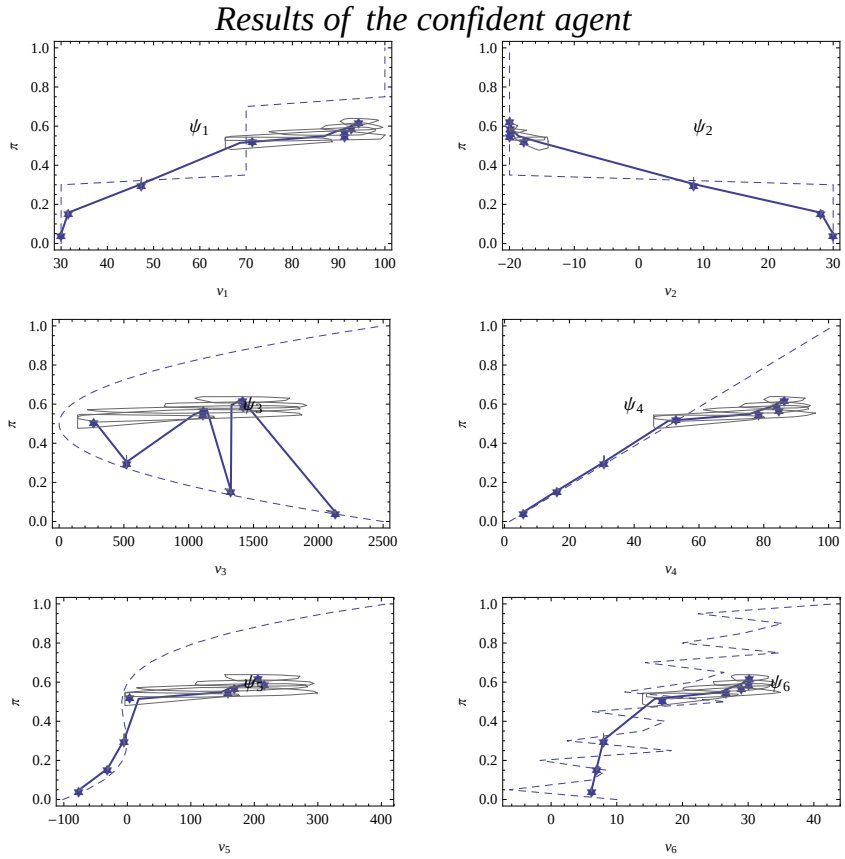


Figure 8:

bounded by 0.75. On the contrary, $\{\varphi_k(x_1), \cdots, \varphi_k(x_{35})\}$ may be quite scattered, and $\overline{\varphi_k}(\overrightarrow{x}) := \epsilon + \frac{1}{35}\sum_{i=1}^{35}\varphi_k(x_i)$, where $\epsilon \approx N(0, 0.1)$ may be rather fluctuating (see Figures 8 & 10).

5.1.2. Using a regular grid

The simulated covers were afterward coarsened by using a regular grid, with the same number of intervals than the GIPREB one (see Figure 11), giving rise to another 30 by 35 matrix M of codes. This grid is much better suited for our simulated data. In addition, remember that in this case the confident and the suspicious agents are confounded (see Appendix B); consequently, we only plotted the pignistic distribution corresponding to the first one, on Figure 12; naturally, it is also rather close to the uniform distribution.

We can see on Figure 13 that the bijective functions (a linear and a cubic one) were very well restored. The injective application ψ_3 was very well restored too, and one can infer from the associated panel that it is not a function of ν_3;

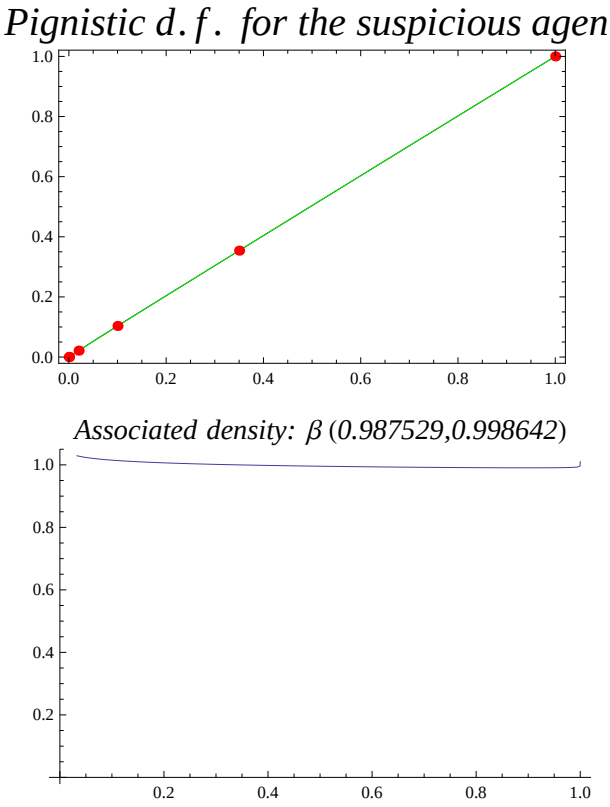


Figure 9:

thus, π cannot be predicted by ν_3. Linking the averages in function of the ordinate instead of the abscissa, the reader could see that, on the contrary, its lower inverse φ_3 is indeed a function of π (see the confidence regions). The restoration of ψ_6 was rather poor, but notice that it was a hard problem, since φ_6 is the sum of a linear trend and a cosine of rather high frequency. At last, neither ψ_1 nor ψ_2 is injective; consequently, using CVRR is dubious in these cases. Nevertheless, the trends corresponding to these applications were satisfactorily restored.

5.2. Application to the P. pectinatus data

First of all, we discarded six stations (30-35) from the data set, because their cover was always null. Theses stations are located is the South part of the lagoon, and are characterized by unfavorable physical (strong currents) and sedimentary (gravel and shells) conditions for *P. pectinatus* growth.

The pignistic probability associated with $\mathcal{B}_C$ is (see Appendix B):

$$BetP_{\mathcal{C}} = (0.2389146, 0.2610854, 0.2610854, 0.2389146),$$

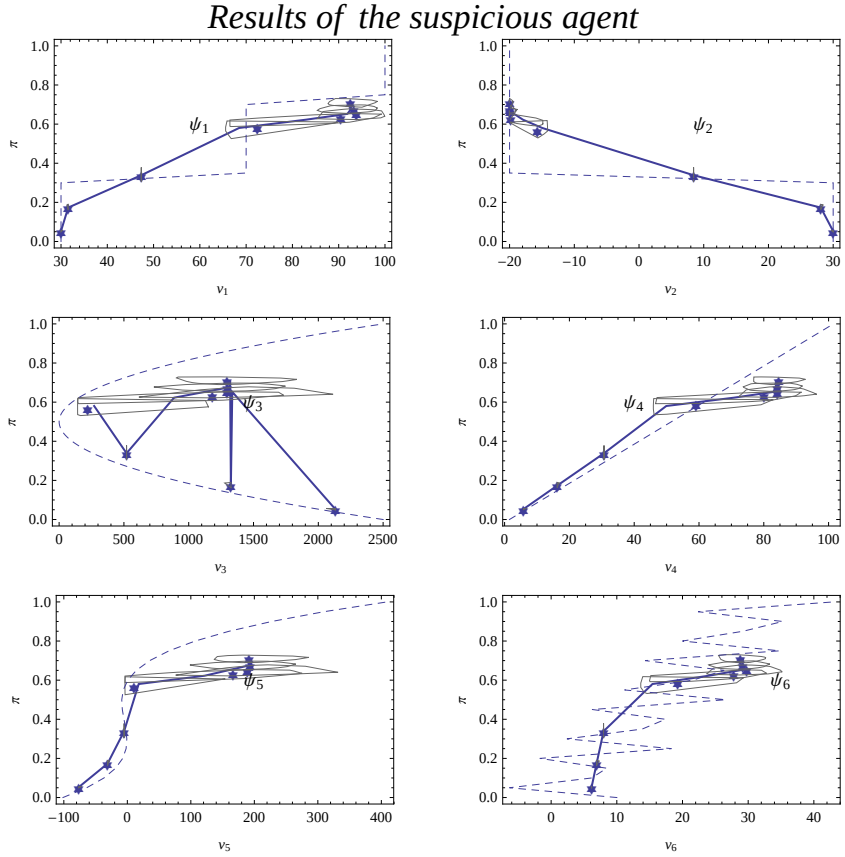


Figure 10:

and Figure 14 shows that the corresponding d.f. is rather well fitted by a beta distribution (the biggest fitting error is: $P(\beta(0.324432, 4.41761) > 0.35) = 0.0329426$); notice that, contrary to previous pignistic probabilities, it is supported by the cover interval associated with Table 1.

Two hundred preliminary independent runs were performed and, using DW as the distance between restored states, the optimal number of classes was 1 (6.5% of runs), 2 (78%) or 3 (15.5%). Consequently, the correct number of groups should not exceed three. We decided to set it to three, because it was the result already obtained with the central score (see Section 3.2.1), and also because with three classes (or more - see Section 5.1) it is possible to reveal nonlinear relationships.

We subsequently performed two hundred other runs, clustering the restored states into three groups, and associating to each run a per group regression function similar to that displayed on Figure 5. We obtained this way 200 triples $\{(\tilde{m}_Z(g_1^s), \tilde{m}_Z(g_2^s), \tilde{m}_Z(g_3^s)) : 1 \leq s \leq 200\}$ for each variable Z (including the

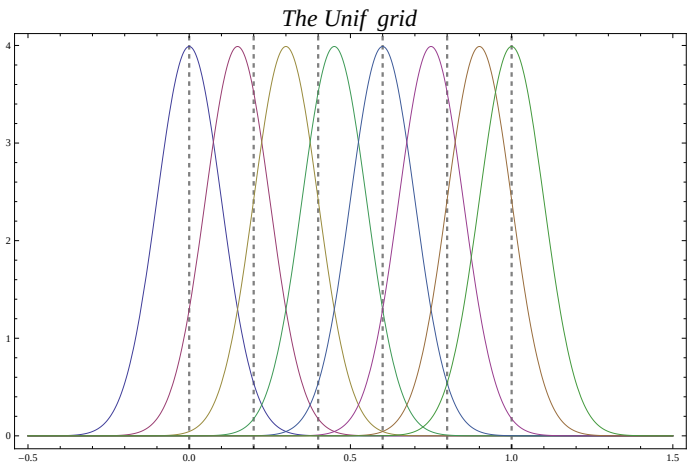


Figure 11:

cover itself). Remember $\tilde{m}_Z\left(g_k^s\right)$ is the average of all the values of Z associated with observations (years) assigned to the kth group, in the sth run. It is also important to note that, for coherency, these random groups of years were labeled in function of the target variable Y (the cover), in order that we have:

$$\tilde{m}_Y\left(g_1^s\right) \leq \tilde{m}_Y\left(g_2^s\right) \leq \tilde{m}_Y\left(g_3^s\right) \quad \forall s,\, 1 \leq s \leq 200.$$

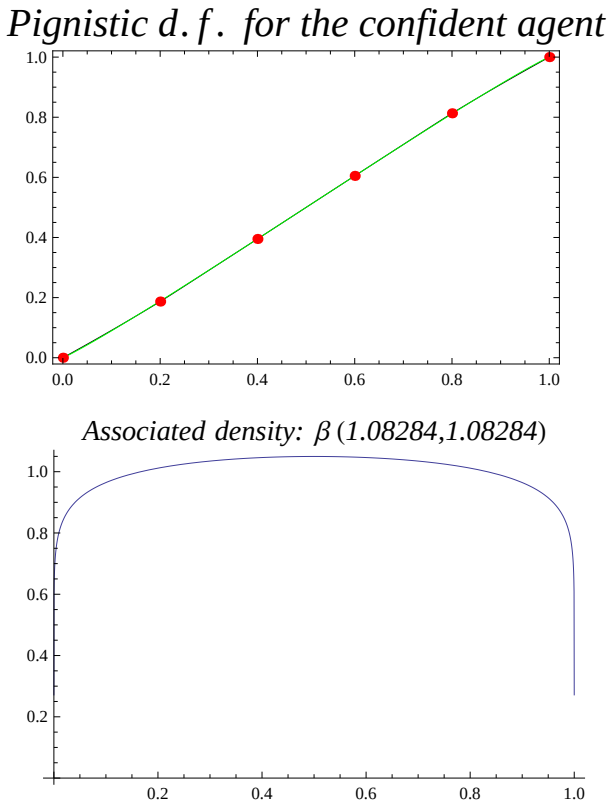
Thanks to this sorting, we are able to associate the groups issued from different runs, and to characterize the relationship between Y and each hydrological variable X by the three conditional distributions associated with the groups (see Figures 8 & 15). On these plots, the points are naturally linked by segments, in function of their abscissas.

A synthesis of the triples obtained from these 200 runs

$$\left\{\left(\tilde{m}_X\left(g_1^s\right), \tilde{m}_Y\left(g_1^s\right)\right), \left(\tilde{m}_X\left(g_2^s\right), \tilde{m}_Y\left(g_2^s\right)\right), \left(\tilde{m}_X\left(g_3^s\right), \tilde{m}_Y\left(g_3^s\right)\right) : 1 \leq s \leq 200\right\}$$

is displayed on Figure 15. Each panel of this figure crosses the restored cover (the ordinate Y) with some hydrological variables (the abscissa X). The structure of the three groups (the associated conditional distributions, summed up by confidence regions, medians and averages) crystallizes the possible relationship between X and the cover. The different panels of Figure 15 highlights the variables which could actually explain the cover of *P. pectinatus*: Sal., Sal.Aug., FW-1, Silt, [N-NO_3] and [P-PO_4]. On the contrary, since the relationships between the cover and the other hydrological variables (FW, N-NO_3, [SS] and

Figure 12:



P-PO_4) are not monotonous, they could not explain it straightforwardly. It is noteworthy that most of the good predictors were retained by CCA; nevertheless, surprisingly, CCA discarded Silt.

But what about the suspicious agent $\mathcal{B}_S$? In his case, the pignistic probability is (see Appendix B):

$$BetP_{\mathcal{S}} = (0.0014277, 0.0558381, 0.2287379, 0.713996),$$

which is close to the uniform distribution on $[0, 0.35]$. Since the results obtained by both of the agents were very similar, we focused only on those of $\mathcal{B}_C$, for sake of brevity.

5.2.1. Ecological interpretation

Trough the whole observation series, among the several hypotheses which could account for the year-to-year variation in the extent of *P. pectinatus* in the Berre lagoon, the inputs of freshwater generating the drop in salinity may have constituted one of the main controlling factors. *P. pectinatus* is a freshwater species; the plant frequently inhabits rivers, streams, coastal ponds and

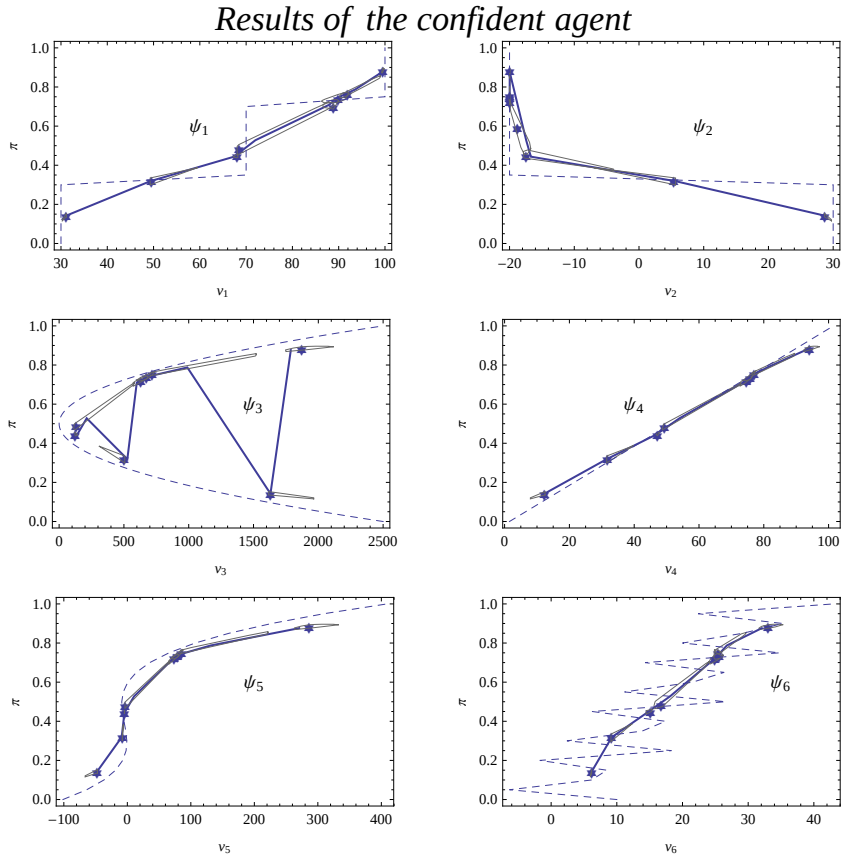


Figure 13:

estuarine wetlands, as it can adapt to moderate salinity and fluctuating water levels (Kantrud, 1990; Pilon et al., 2002). Indeed, we can see on Figure 15 that mean annual salinity (Sal.) and mean salinity in August (Sal.Aug.) are linked with *P. pectinatus* abundance. Of interest is the strong relationship between the cover and total annual inputs of freshwater by the Durance River from October to September (FW-1), which integrates the 12 months-period before the peak-vegetation of *P. pectinatus*. On the contrary, the annual input summed from January to December (FW) cannot predict the abundance of this species. Inputs of silts from the Durance River are linked negatively with the *P. pectinatus* abundance. Water transparency frequently limits the depth distribution of aquatic angiosperms (Bowen and Valiela, 2001; Valiela et al., 1997) and the heavy inputs of silts up to the latter 1970s could have constrained the *P. pectinatus* growth, although this species is well-adapted to turbid environment (Kantrud, 1990). Eventually, the reduction of silts loads in the early 1980s may have increase the water transparency and, as a consequence, improve the growth

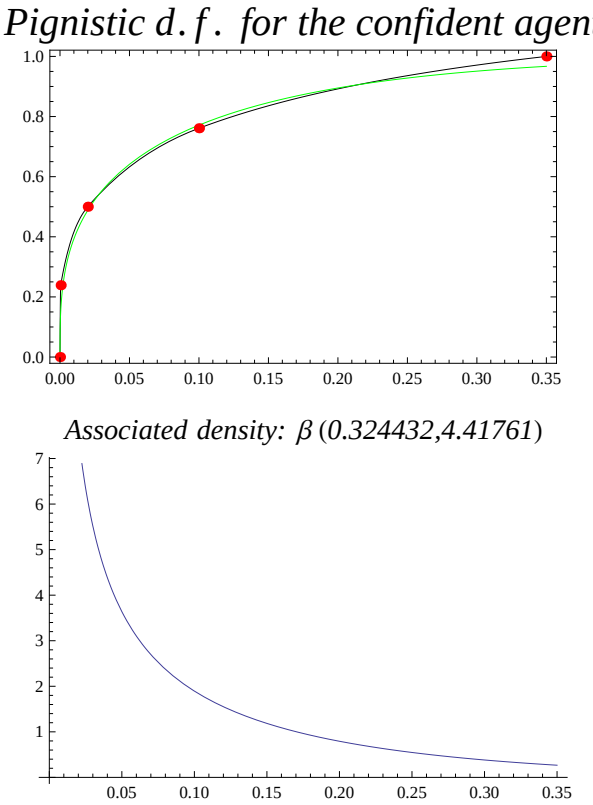


Figure 14:

of *P. pectinatus*. Finally, *P. pectinatus* abundance is linked with concentrations in phosphate and nitrate, while that is not the case with inputs of these chemical elements. In surface water, concentrations in phosphates remained low through the whole time series, close to the value given to be limiting for *P. pectinatus* growth. This can cope with the common assumption that phosphorus is usually the limiting factor for growth of freshwater macroalgae (Van Wijk, 1989).

6. Conclusion

We have proposed an original method for explaining a vector of imprecise interval data with several quantitative variables. It combines stochastic restoration (based on the Transferable Belief Model), clustering and nonparametric regression. After a preliminary test on realistic simulated data, it has been applied to explain the annual cover variations of *Potamogeton pectinatus* in the Berre lagoon from 1970 to 2004 with ten hydrological variables. Our results suggest that freshwater inputs, salinity and nutrient were the forcing variables responsible for the phase shifts from *P. pectinatus* meadows to bare silt habitat.

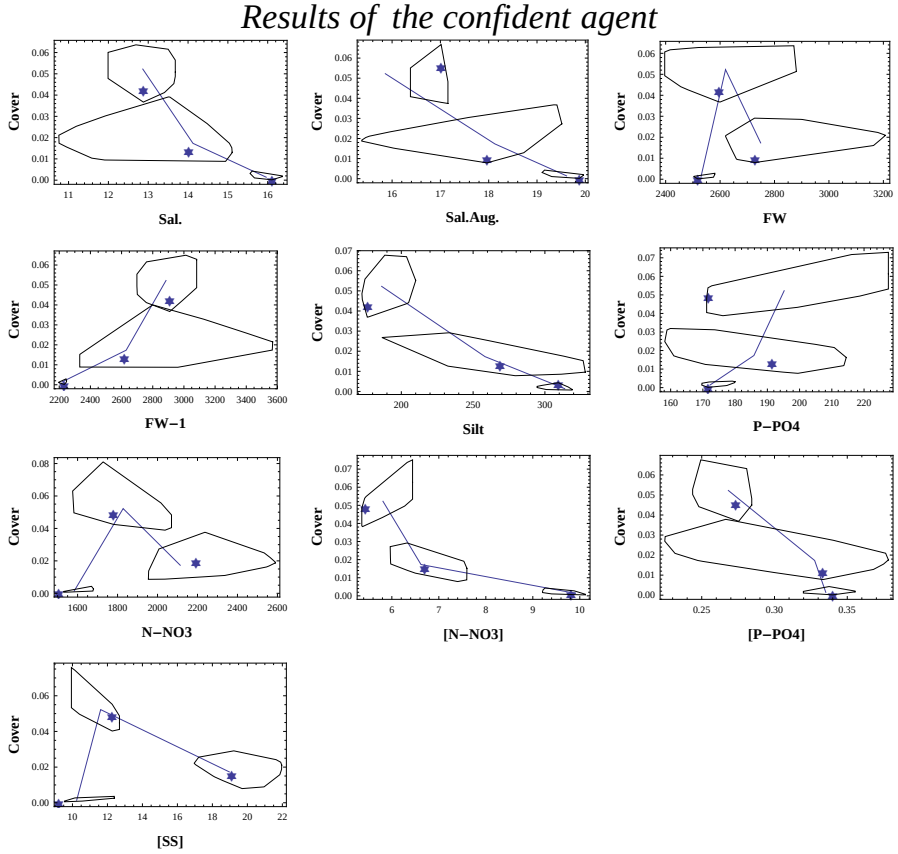


Figure 15:

Through the whole time series, the *P. pectinatus* extent has been enhanced by respectively increasing freshwater inputs and decreasing salinity, while it has been limited by phosphorus and nitrogen concentrations in water.

The obtained issues were basically in accordance with those from CCA, but notice that CVRR can deal with nonlinear relationships. Moreover, CCA does not take into account the nature of such data, contrarily to CVRR. Furthermore, the proposed restoration method takes into account hypotheses about the diver's behavior. Of course, other hypothetical behaviors could be tried. It could also be possible to obtain pignistic probabilities from the concrete experience of divers.

We considered the ten explanatory variables one at a time, plotting characteristics of the clusters. The per group regression consisted in linking the obtained average conditional expectations:

$$\left\{\left(\widetilde{m}_{\pi}\left(g_1\right),\overline{\widetilde{m}_{Y}}\left(g_1\right)\right),\left(\widetilde{m}_{\pi}\left(g_2\right),\overline{\widetilde{m}_{Y}}\left(g_2\right)\right),\left(\widetilde{m}_{\pi}\left(g_{13}\right),\overline{\widetilde{m}_{Y}}\left(g_3\right)\right)\right\}$$

where Y denotes any explanatory variable. It would be conceivable to su-

persede these separate simple regressions by a single multiple regression, if we had much more than three clusters! This remark points out that CCVR has been described as an "inter-groups" method. It could be easily completed by "intra-groups" analyzes (simples or multiples regressions for instance). This would not be very interesting here, because of the small size of our data set (12 observations, necessarily duplicated during the restoration process).

In conclusion, the results obtained on real data as well as on realistic (but richer!) simulated data show that CVRR can reveal nonlinear relationships between the coarsened cover and hydrological variables, whichever these relationships are one-to-one. Consequently, it should be generally helpful in Ecology (or other disciplines) for investigating the relationships between vectors of coarse responses and explicative variables.

Acknowledgements

The data on inflow of freshwater and silt due to the diversion of the Durance river towards the Berre lagoon, via the Saint-Chamas hydroelectric plant (1966 through 2004), has been provided by EDF (Electricité de France). Data on the inflow of freshwater due to rivers flowing into the lagoon, on the inflow of nutrients due to either the diversion of the Durance River or to rivers flowing into the lagoon, and on nutrient concentrations in the water come from the literature (Minas, 1974; Arfi, 1989; Kim and Travers, 1997b; Marcos-Diego et al., 2000; Gouze et al., 2008), and from the French Ministry of Environment databases (EauFrance, 2006). Salinity (1994 to 2004) was measured with a CTD probe YSI®, every 50 cm down to 4 m. The authors are very grateful to E. Tapia Moore for help with the English text.

Appendix A: the TBM model for a single sensor

The TBM has been widely used to model fusion of expert opinions (Ha-Duong, 2008) and fusion of sensors data (Delmotte and Smets, 2004), amongst other topics. Particularily, the "one sensor problem" handled by Delmotte and Smets (2004) has been a source of inspiration for us. These authors suppose there is a set of n targets $C := \{c_1, \cdots, c_n\}$ (various aircrafts, for instance); the actual (unknown) one is $\mathring{c} \in C$, while the set of possible observations (images, for instance) is Π. The identification problem consists in estimating $\mathring{c}$ from a set $\pi^k := \{\pi_1, \cdots, \pi_k\}$ of independent observations of the same target $\mathring{\pi}$. Delmotte and Smets (2004) proposed to estimate $\mathring{c}$ by $\hat{c} := \arg\max_{c \in C} BetP\left[\mathring{\pi}\right](c)$, where $BetP\left[\mathring{\pi}\right]$ denote the pignistic probability associated with the observation of $\mathring{\pi}$. Concretely (Smets, 2005b), $BetP\left[\mathring{\pi}\right]$ is computed from a "basic belief assignment" (**bba**) $m\left[\mathring{\pi}\right](A)$ allocating to any logical proposition A about $\mathring{c}$ some amount of belief (in $[0,1]$), conditionally to the observation of $\mathring{\pi}$. Delmotte and Smets (2004) proved that in the "one sensor problem" case, $m\left[\mathring{\pi}\right]$ can be computed from the conditional probability measures $\{P(\pi|c_1), \cdots, P(\pi|c_n)\}$ on Π. Thus, this method requires knowledge of these probabilities, obtained for

instance from a confusion matrix built by training the sensor in a preliminary step.

Appendix B: determining pignistic probabilities from symbolical belief functions

In the TBM, a pignistic probability $BetP$ results from the "pignistic transformation" of a given basic belief assignment (bba) m (Smets, 1999, 2005b), in one-to-one correspondence with a belief function Bel. There is a number of belief functions (Dubois et al., 2008) corresponding to a unique pignistic probability $BetP$; to obtain a unique bba m from $BetP$, it is necessary to impose constraints to m.

Aregui and Denoeux (2008) constructed this way a belief function from a pignistic probability; we will follow the opposite way, building pignistic probabilities from symbolical belief functions. Each belief function will be associated with the evidential corpus (or "frame of discernment" (Shafer, 1976)) of some "agent" (alias $\mathcal{B}$). The evidential corpus of $\mathcal{B}$ consists of logical propositions, corresponding to "all $\mathcal{B}$ knows" (Smets, 1998; Smets and Kennes, 1994). We will build belief functions corresponding to two agents, with different behaviours: a confident one ($\mathcal{B}_C$), and a suspicious one ($\mathcal{B}_S$).

6.1. The nature of Bel and m, and their relationships

Let Ω be a finite set of logical propositions.

Definition 1. (Shafer, 1976) A belief function Bel is a function from the power set $\mathcal{P}(\Omega)$ (frequently denoted 2^{Ω}) to $[0,1]$ such that, for any family $\{A_1, \cdots, A_n\}$ of elements of $\mathcal{P}(\Omega)$

$$Bel(\emptyset) = 0 \tag{1}$$

$$Bel\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} Bel\left(A_i\right) - \sum_{i>j} Bel\left(A_i \cap A_j\right) \cdots + (-1)^{n+1} Bel\left(\bigcap_{i=1}^{n} A_i\right). \tag{2}$$

The associated basic belief assignment m, is the Moebius representation of Bel, defined by Shafer (1976):

$$\forall A \in \mathcal{P}(\Omega),\ m\left(A\right) := \sum_{\emptyset \neq B \subseteq A} (-1)^{|A-B|} Bel(B) \tag{3}$$

where $|E|$ denotes the cardinality of E. This function fulfills the conditions

$$\sum_{A \in \mathcal{P}(\Omega)} m\left(A\right) = 1$$
$$m\left(\emptyset\right) = 1 - Bel\left(\Omega\right).$$

Reciprocally, we have the relation:

$$\forall A \in \mathcal{P}(\Omega),\ Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B).$$

In addition, we will suppose that m is normal, *i.e.* $m(\emptyset) = 0$ (this "closed-world" assumption means that the frame of discernment certainly contents the true value of the variable of interest(Smets, 1990)- this is clearly our case). Let now $\hat{\pi}$ be some observation by $\mathcal{B}$, and let $G \in \mathcal{P}(\Omega)$. According to Smets (1990, 1999), $Bel(G)$ is the **total** amount of belief of $\mathcal{B}$ supporting the proposition "$\hat{\pi} \in G$", while $m(G)$ is the **partial** amount of belief of $\mathcal{B}$ supporting this proposition, which does not support any strict subset of G (due to lack of precision, insufficient information, *etc*). For instance, $\mathcal{B}$ believes that $\hat{\pi} \in A \cup B$, but is unable to choose between $\hat{\pi} \in A$ or $\hat{\pi} \in B$.

6.2. The pignistic probability associated with a normal basic belief assignment

The pignistic transformation (Smets, 1999, 2005b) makes possible to skip from the credal level to the pignistic one, in order to take decisions.

Definition 2. The pignistic probability distribution associated with m is $BetP$, defined on Ω as

$$\forall \omega \in \Omega,\ BetP(\omega) := \sum_{A \subseteq \Omega | \omega \in A} \frac{m(A)}{|A|}.$$

Remark 3. An interesting characteristic of $BetP$ is that it is the centre of gravity of the set of probabilities dominating the belief function Bel (Dubois et al., 2008).

6.3. Calculating m and BetP in the case of a family of contiguous intervals

Suppose ω_0 is an observation of $\mathcal{B}$, and consider the algebra generated by the finite set of propositions

$$\Omega := \{\omega_0 \in [a_1, a_2[\,, \omega_0 \in [a_2, a_3[\,, \cdots, \omega_0 \in [a_I, a_{I+1}]\} := \{C_1, \cdots, C_I\}.$$

For simplicity of notations, we will identify each C_i with the associated interval, and suppose $a_1 = 0$ and $a_{I+1} = 1$.

First, what is the meaning of the number $Bel(C_i)$ in this work? Suppose $\mathcal{B}$ observed π, detected $\hat{\pi}$ and decided $O(\hat{\pi}) = i \in \{0, 1, 2, 3, 4\}$. Then, $Bel(C_i)$ measures the belief of $\mathcal{B}$ in the proposition $\pi \in C_i$; in other words, the function Bel measures the confidence of $\mathcal{B}$ in the apparatus used (including $\mathcal{B}$ himself).

Now, because of formula (3) $\forall i$, $m(C_i) = Bel(C_i)$ and, for any pair of intervals

$$Bel(C_i \cup C_j) = Bel(C_i) + Bel(C_j) + m(C_i \cup C_j) \geq 0. \tag{4}$$

Thus, $m(C_i \cup C_j) = Bel(C_i \cup C_j) - (Bel(C_i) + Bel(C_j))$ is a determination of the indiscernibility of C_i and C_j for $\mathcal{B}$. Naturally, this quantity should be a

decreasing function of the size of the void between the intervals. Consequently, we should have

$$m\left(C_i \cup C_{i\pm1}\right) \geq m\left(C_i \cup C_{i\pm2}\right) \geq m\left(C_i \cup C_{i\pm3}\right), \; etc. \tag{5}$$

and, for k great enough, $m\left(C_i \cup C_{i\pm k}\right)$ should be very small.

6.3.1. $\mathcal{B}_C$, a confident agent

Suppose $\mathcal{B}_C$ highly trusts the coding used, *i.e.* he thinks the atoms of Ω consist of intervals perfectly suited for coding the capability of observation of a diver. Then the same credal level must be assigned to each interval: $\forall i,\, Bel\left(C_i\right) = \mu$. Notice that this implies $\mu \leq \frac{1}{I}$, since $1 = Bel\left(\Omega\right) = Bel\left(\bigcup_{i=1}^{I} C_i\right) \geq \sum_{i=1}^{I} Bel\left(C_i\right) = I\,\mu$. Since all the intervals possess the same degree of belief, we can also reasonably postulate that $\forall i,\, m\left(C_i \cup C_{i\pm1}\right) = \varepsilon \in]0,1[$.

Introduce now two auxiliary functions, Δ and δ defined on the complement $\mathcal{P} \ominus \mathcal{S}$ of the singletons in the power set $\mathcal{P}(\{1,\cdots,I\})$:

$$\mathcal{P} \ominus \mathcal{S}(\{1,\cdots,K\}) := \mathcal{P}(\{1,\cdots,K\}) - \{\{1\},\cdots,\{K\}\}).$$

These functions are $\Delta(E) := \max_{i,j\in E} |i-j|$ and $\delta(E) := \min_{i,j\in E} |i-j|$. In accordance with inequality (5), we postulate in addition that

$$\forall E \in \mathcal{P} \ominus \mathcal{S}(\{1,\cdots,I\}),\; m\left(\bigcup_{i\in E} C_i\right) = \varepsilon^{\Delta(E)}.$$

Then, we can write:

$$\begin{aligned}\forall i,\, BetP\left(C_i\right) &= m\left(C_i\right) + \tfrac{1}{2}\sum_{j=i\pm k} m\left(C_i \cup C_j\right) + \tfrac{1}{3}\sum_{(i,j,k)|\delta(\{i,j,k\})>0} m\left(C_i \cup C_j \cup C_k\right) + \cdots \\ &= \mu + \tfrac{1}{2}\sum_{j=i\pm k} \varepsilon^{\Delta(\{i,j\})} + \tfrac{1}{3}\sum_{(i,j,k)|\delta(\{i,j,k\})>0} \varepsilon^{\Delta(\{i,j,k\})} + \cdots\end{aligned}$$

These probabilities can be easily calculated in the case of the *Potamogeton* data ($I = 4$):

$$\begin{aligned}\Xi\, BetP\left(C_1\right) &= \mu + \tfrac{1}{2}\left(\varepsilon + \varepsilon^2 + \varepsilon^3\right) + \tfrac{1}{3}\left(\varepsilon^2 + 2\,\varepsilon^3\right) + \tfrac{1}{4}\varepsilon^3 = \mu + \tfrac{1}{2}\varepsilon + \tfrac{5}{6}\varepsilon^2 + \tfrac{17}{12}\varepsilon^3 \\ \Xi\, BetP\left(C_2\right) &= \mu + \tfrac{1}{2}\left(\varepsilon^2 + 2\,\varepsilon\right) + \tfrac{1}{3}\left(\varepsilon^2 + \varepsilon^3 + \varepsilon^2\right) + \tfrac{1}{4}\varepsilon^3 = \mu + \varepsilon + \tfrac{7}{6}\varepsilon^2 + \tfrac{7}{12}\varepsilon^3 \\ \Xi\, BetP\left(C_3\right) &= \mu + \tfrac{1}{2}\left(\varepsilon^2 + 2\,\varepsilon\right) + \tfrac{1}{3}\left(\varepsilon^2 + \varepsilon^3 + \varepsilon^2\right) + \tfrac{1}{4}\varepsilon^3 = \mu + \varepsilon + \tfrac{7}{6}\varepsilon^2 + \tfrac{7}{12}\varepsilon^3 \\ \Xi\, BetP\left(C_4\right) &= \mu + \tfrac{1}{2}\left(\varepsilon + \varepsilon^2 + \varepsilon^3\right) + \tfrac{1}{3}\left(\varepsilon^2 + 2\,\varepsilon^3\right) + \tfrac{1}{4}\varepsilon^3 = \mu + \tfrac{1}{2}\varepsilon + \tfrac{5}{6}\varepsilon^2 + \tfrac{17}{12}\varepsilon^3\end{aligned}$$

where the total mass is $\Xi = 4\,\mu + 3\,\varepsilon + 4\,\varepsilon^2 + 4\,\varepsilon^3$.

Fixing $\mu = \frac{1}{4}$ and $\varepsilon = 0.05$ (which corresponds to an admissible degree of indiscernibility: 20% of the basic degree of belief μ), we obtain the pignistic probability $BetP_C = (0.238915, 0.261085, 0.261085, 0.238915)$.

6.3.2. $\mathcal{B}_S$, a not so much confident agent

Consider now our second agent, $\mathcal{B}_S$, and let us denote $\lambda(G)$ the Lebesgue measure of some interval G. $\mathcal{B}_S$ does not trust much the imposed system of intervals, and allocates to each atom a credal level proportional to its width: $\forall i,\, m(C_i) = Bel(C_i) = \eta_i := \alpha\,(a_{i+1} - a_i)$; since Ω is the unit interval, $Bel(\Omega) = 1 = \lambda(\Omega) \Rightarrow \alpha = 1$.

Consider now the indiscernibility determination $m(C_i \cup C_{i\pm1}) = Bel(C_i \cup C_{i\pm1}) - (\eta_i + \eta_{i\pm1})$. We postulate that the difficulty for $\mathcal{B}_S$ of choosing between C_i and $C_{i\pm1}$ should depend on the ratios $\left(\frac{\eta_i}{\eta_{i\pm1}}, \frac{\eta_{i\pm1}}{\eta_i}\right)$ on the one hand (their balance), and on the size of $C_i \cup C_{i\pm1}$, $\eta_i + \eta_{i\pm1}$, on the other hand. The influence of the balance between intervals will be quantified by the arbitrary function

$$Q\left(x_{\{i,i\pm1\}}\right) := \left(1 - \min\left(\left(\frac{1}{x_{\{i,i\pm1\}}} - 1\right)^2, \left(x_{\{i,i\pm1\}} - 1\right)^2\right)\right)^5,$$

where $x_{\{i,i\pm1\}} := \min\left(\frac{\eta_i}{\eta_{i\pm1}}, \frac{\eta_{i\pm1}}{\eta_i}\right) \in]0, 1]$. The function $Q(x)$ is represented on Figure 16; it is maximum and equals 1 at $x = 1$ (balanced intervals), and practically vanishes for $x < 0.2$.

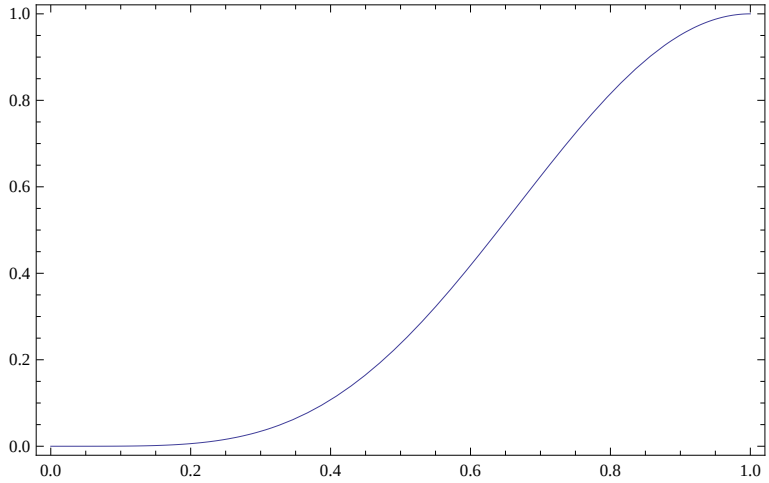


Figure 16:

In addition, we will impose for consistency that **when all the atoms are of equal length, $\mathcal{B}_S$ would take the same decision as $\mathcal{B}_C$**. Consequently,

we postulate for contiguous intervals:

$$\Xi\, m\left(C_i \cup C_{i\pm1}\right) := \frac{2\varepsilon}{I}\, \frac{Q\left(x_{\{i,i\pm1\}}\right)}{\eta_i + \eta_{i\pm1}}$$

(where $\Xi := \sum_{A\in\mathcal{P}(\Omega)} m\left(A\right)$ denotes the total *a posteriori* mass), because the indiscernibility should naturally be a decreasing function of $\lambda\left(C_i \cup C_{i\pm1}\right)$.

As for pairs of non-contiguous intervals, we will also take into account the void:

$$\begin{aligned}
\Xi\, m\left(C_i \cup C_{i+k}\right) &:= \tfrac{2\varepsilon^k}{I-k+1}\, \tfrac{Q\left(x_{\{i,i\pm k\}}\right)}{\eta_i+\eta_{i+k}} \left(1 - \textstyle\sum_{j=1}^{k-1} \eta_{i+j}\right) \\
\Xi\, m\left(C_i \cup C_{i-k}\right) &:= \tfrac{2\varepsilon^k}{I-k+1}\, \tfrac{Q\left(x_{\{i,i\pm k\}}\right)}{\eta_i+\eta_{i-k}} \left(1 - \textstyle\sum_{j=1}^{k-1} \eta_{i-j}\right)
\end{aligned}$$

since $m\left(C_i \cup C_{i+k}\right)$ should naturally decrease when the void size increases. Consider now the sets $E := \{i, i+k_1, \cdots, i+k_{L-1}\}$ and $\bigcup_{e\in E} C_e$, where $0 = k_0 < k_1 < \cdots < k_{L-1} < I - i$. Denoting $S_E := \sum_{l=0}^{L-1} \eta_{i+k_l}$ the mass of the support and $V_E := \sum_{j=0}^{k_{L-1}} \eta_{i+j} - S_E$ the mass of the voids, we define the indiscernibility of order L (L intervals involved) by:

$$\Xi\, m\left(\bigcup_{e\in E} C_e\right) := \varepsilon^{\Delta(E)} Q\left(x_E\right)\, \frac{L}{I+L-1-k_{L-1}}\, \frac{(1-V_E)}{S_E}$$

where $x_E := \max_{i,j\in E} \min \frac{\eta_i}{\eta_j}$ measures the global balance of the supporting intervals.

Fixing again $\varepsilon = 0.05$, we obtain as pignistic probability in the case of the *Potamogeton* data

$$BetP_S = \left(0.00142514, 0.0562086, 0.229236, 0.71313\right).$$

Figures captions

Figure 1: Study site and sampling stations (indicated by numbers 1 to 35).

Figure 2: Total abundance index of *P. pectinatus* in the Berre lagoon from 1970 through 2004 (points). Groups of years (according to the metric DW) are indicated by ellipses. Vertical lines indicates the "shift" between state (a) and state (b) (Rodionov, 2004: shift of the mean; probability = 0.1; length cut-off=4 years; Huber parameter = 1).

Figure 3: Linear model obtained between T and [N-NO_3].

Figure 4: First principal plane of CCA. Upper panel: hydrological variables; lower panel: observations and cover codes.

Figure 5: A per group regression function: the abscissas and ordinates of these three points are conditional expectations per group of observations (years).

Figure 6: The completed GIPREB grid (vertical lines), and the distributions $\{\mathcal{D}_1, \cdots, \mathcal{D}_8\}$.

Figure 7: Completed GIPREB grid; distribution function of the pignistic probability, superimposed to the associated Beta distribution, in green. In black (but nearly hidden): the Stineman interpolation.

Figure 8: Completed GIPREB grid; recovering of relationships between a coarsened vector of simulated "cover" (eight clusters) and different types of "hydrological variables", corrupted by additive errors. On each plot, each group is represented by three statistical characteristics: the convex polygon containing 75% of the data; the bivariate median (labeled by a star), and the average. Averages are linked by segments according to the sorted abscissas, giving rise to per group regressions; the true functions are dashed.

Figure 9: Completed GIPREB grid; distribution function of the pignistic probability, superimposed to the associated Beta distribution, in green. The Stineman interpolation is hidden by the Beta d.f.

Figure 10: Completed GIPREB grid; recovering of relationships between a coarsened vector of simulated cover (eight clusters) and different types of hydrological variables, corrupted by additive errors. On each plot, each group is represented by three statistical characteristics: the convex polygon containing 75% of the data; the bivariate median (labeled by a star), and the average. Averages are linked by segments according to the sorted abscissas, giving rise to per group regressions; the true functions are dashed.

Figure 11: The grid used (vertical lines), and the distributions $\{\mathcal{D}_1, \cdots, \mathcal{D}_8\}$.

Figure 12: Regular grid; distribution function of the pignistic probability, superimposed to the associated Beta distribution, in green. The Stineman interpolation is hidden by the Beta d.f.

Figure 13: Regular grid; recovering of relationships between a coarsened vector of simulated "cover" (eight clusters) and different types of "hydrological variables", corrupted by additive errors. On each plot, each group is represented by three statistical characteristics: the convex polygon containing 75% of the data; the bivariate median (labeled by a star), and the average. Averages are linked by segments according to the sorted abscissas, giving rise to per group regressions; the true functions are dashed.

Figure 14: Original GIPREB grid; distribution function of the pignistic probability, superimposed to the associated Beta distribution, in green. In black: the Stineman interpolation.

Figure 15: Statistical characteristics of the per group regression function. On each plot, each group is represented by three statistical characteristics: the convex polygon containing 75% of the restored data; the bivariate median (labeled by a star), and the average. Averages are linked by segments according to the sorted abscissas, giving rise to per group regression functions.

Figure 16: The function $Q(x)$.

References

Aregui, A. and Denoeux, T. (2008). Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. *International Journal of Approximate Reasoning*, **49**, 575–594.

Arfi, R. (1989). Annual cycles and budget of nutrients in Berre lagoon (Mediterranean, France). *Internationale Revue der Gesamten Hydrobiologie*, **74**, 29–49.

Berge, C. (1966). *Espaces topologiques. Fonctions multivoques*. Dunod, Paris.

Bernard, G., Bonhomme, P., and Boudouresque, C. F. (2005). Recovery of the seagrass *Zostera marine* in a disturbed Mediterranean lagoon (Etang de Berre, Southern france). *Hydrobiologia*, **539**, 157–161.

Bernard, G., Boudouresque, C. F., and Picon, P. (2007). Long term changes of *Zostera* meadows in the Berre lagoon (Provence, Southern France). *Estuar. Coast Shelf S.*, **73**, 617–629.

Billard, L. and Diday, E. (2006). *Symbolic data analysis: conceptual statistics and data mining*. Series on Computational Statistics. Wiley, London.

Bowen, J. and Valiela, I. (2001). The ecological effects of urbanization of coastal watersheds: Historical increases in nitrogen loads and eutrophisation of Waquoit Bay Estuaries. *Can. J. Fish. Aquat. Sci.*, **58**, 1489–1500.

Brock, W. and Carpenter, S. (2006). Variance as a leading indicator of regime shift in ecosystem service. *Ecol. and Soc.*, **11**, **2**, 9–24.

Chen, J., Shiyomi, M., Bonham, C. D., Yasuda, T., Yoshimichi, H., and Yamamura, Y. (2008a). Plant cover estimation based on the Beta distribution in grassland vegetation. *Ecol. Res.*, **23**, 813–819.

Chen, J., Shiyomi, M., Yoshimichi, H., and Yamamuta, Y. (2008b). Frequency distribution models for spatial patterns of vegetation abundance. *Ecol. Model.*, **11**, 403–410.

De Leuw, J. and Mair, P. (2009). Simple and Canonical Correspondence Analysis using the `R package anacor`. *J. Stat. Softw.*, **31**, **53**, 1–18.

Delmotte, F. and Smets, P. (2004). Target identification based on the Transferable Belief Model interpretation of Dempster-Shafer model. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, **34**, **4**, 457 - 471.

Dubois, D., Prade, H., and Smets, P. (2008). A definition of subjective possibility. *International Journal of Approximate Reasoning*, **48**, 352–364.

EauFrance (2006). Banque hydro. http://www.hydro.eaufrance.fr.

Einbeck, J. and Tutz, G. (2006a). The fitting of multifunctions: an approach to nonparametric multimodal regression. *COMPSTAT: Proceedings in Computational Statistics, 17th Symposium, Roma, Italy*, pages 1251–1258.

Einbeck, J. and Tutz, G. (2006b). Modelling beyond regression functions: an application of multimodal regression to speed-flow data. *Applied Statistics*, **55**, **4**, 461-475.

Gibbs, A. L. and Su, F. O. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.*, **70**, 419–435.

Gouze, E., Raimbault, P., Garcia, N., Bernard, G., and Picon, P. (2008). Nutrient and suspended matter discharge by tributaries into the Berre lagoon (France): the contribution of flood event to the matter budget. *Geoscience*, **340**, 233–244.

Greenacre, M. (2007). *Correspondence Analysis in Practice*. CRC Interdisciplinary Statistics. Chapman and Hall, Boca Raton.

Guizan, A. and Harrell, F. (2000). Ordinal response regression models in ecology. *J. Veg. Sci.*, **11**, 617–626.

Ha-Duong, M. (2008). Hierarchical fusion of expert opinions in the transferable belief model, application to climate sensitivity. *International Journal of Approximate Reasoning*, **49**, 555–574.

Halmos, P. (1997). *Introduction á la théorie des ensembles*. Jacques Gabay, Paris.

Heitjan, D. F. and Rubin, D. (1991). Ignorability and coarse data. *Ann. Stat.*, **19**, **4**, 2244–2253.

Irvine, K. M. and Rodhouse, T. J. (2010). Power analysis for trend in ordinal cover classes: implications for long-term vegetation monitoring. *J. Veg. Sci.*, **21**, 1152–1161.

Kantrud, H. (1990). Sago pondweed (*Potamogeton pectinatus L.*) a literature review. *U.S. Fish ani Wildlife Service, Fish and Wildlife resource Publication,Jamestown*, **176**, 1–89.

Kaufman, L. and Rousseew, P. (1990). *Finding groups in data: an introduction to Cluster Analysis*. John Wiley and Sons, New York.

Kim, K. and Travers, M. (1997a). Les nutriments de l'Etang de Berre et les milieux aquatiques contigus (eaux douces, saumâtres et marines; Méditerranée NW). 2 Les nitrates. *Marine Nature*, **5**, 35–48.

Kim, K. and Travers, M. (1997b). Les nutriments de l'Etang de Berre et les milieux aquatiques contigus (eaux douces, saumâtres et marines; Méditerranée NW). 4 Les nitrites. *Marine Nature*, **5**, 65–78.

Marcos-Diego, C., Bernard, G., Garcia-Charton, J., and Perez-Ruzafa (2000). Methods for studying impact on *posodonia oceanica* meadows. In *Introductory guide to methods for selected ecological* (eds Goni, R., Harmelin-Vivien, M., Badalamenti, F., Le Dirach, L., and Bernard, G.), pp. 1–120. GIS Posidonie.

Minas, M. (1974). Distribution, circulation et évolution des éléments nutritifs, en particulier du phosphore minéral, dans l'Etang de Berre. Influence des eaux duranciennes. *Internationale Revue der Gesamten Hydrobiologia*, **59**, 509–542.

Mossé, R. and Mossé, J. (1985). Cartographie des peuplements a *Potamogeton pectinatus* dans l'étang de Berre (Bouches-du-Rhône, France). *Commissions Internationales pour la Mer Méditerranée*, **29**, **4**, 123–124.

Nérini, D., Durbec, J., and Manté, C. (2000). Analysis of oxygen rate time series in a strongly polluted lagoon using a regression tree method. *Ecol. Model.*, **133**, 95–105.

Paisley, L. G. and Hostrup-Pedersen, J. (2004). A quantitative assessment of the risk of transmission of bovine spongiform encephalopathy by tallow-based calf milk replacer. *Preventive Veterinary Medicine*, **63**, 135–149.

Pellegrino, R. and Costantino, N. (2012). A Monte Carlo simulation and fuzzy delphi-based approach to valuing real options in engineering fields. In *Risk Management for the Future - theory and cases* (ed. Emblemsvag, J.), pp. 185–214. InTech.

Pilon, J., Santamaria, L., Hootsmans, M., and Van Vierssen, W. (2002). Latitudinal variation in life-cycle characteristics of *Potamogeton pectinatus* L.: vegetative growth and asexual reproduction. *Plant Biol.*, **165**, **2**, 147–262.

Podani, J. (2005). Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. *J. Veg. Sci.*, **16**, 497–510.

Podani, J. (2007). Spatial confusion or clarity ? Reply to Ricotta & Avena. *J. Veg. Sci.*, **18**, 921–924.

Ricotta, C. and Avena, G. (2006). On the evaluation of ordinal data with conventional multivariate procedures. *J. Veg. Sci.*, **17**, 839–842.

Riouall, R. (1972). *Contribution a l'étude de la flore des étangs de Berre et de Vaine (Bouches-du-Rhône)*. PhD thesis, Université d'Aix-Marseille, Marseille, France.

Rodionov, S. (2004). A sequential algorithm for testing climate regime shifts. *Geophys. Res. Lett.*, **31**, doi:10.1029/2004GL019448.

Rodionov, S. and Overland, J. (2005). Application of a sequential regime shift detection method to the Bering sea ecosystem. *ICES Journal of Marine Science*, **62**, 328–334.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, New Jersey.

Singer, J., Poleto, F., and Rosa, P. (2004). Parametric and nonparametric analyzes of repeated ordinal categorical data. *Biometrical J.*, **46**, 460–473.

Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Trans. PAMI*, **12**, 447–458.

Smets, P. (1998). Probability, possibility, belief: which and where? In *Handbook of Defeasible Reasoning and Uncertainty Management Systems* (Gabay, D. and Smets, P., eds.) , vol. 1, pp. 1–24. Kluwer Academic Publishers.

Smets, P. (1999). Practical use of belief functions. In *Uncertainty in Artificial Intelligence* (Laskey, K. B. and Prade, H., eds.), vol. 15, pp. 612–621.

Smets, P. (2005a). Belief functions on real numbers. *International Journal of Approximate Reasoning*, **40**, 181–223.

Smets, P. (2005b). Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, **38**, 133–147.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, **66**, 191–234.

Stora, G. (1976). Evolution des peuplements benthiques d'un étang marin soumis un effluent d'eaux douces. *B. Ecol.*, **7**, **3**, 275–281.

Ter Braak, C. (2006). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradients. *Ecology*, **67**, 1167–1179.

Ter Braak, C. and Venderschot, P. (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, **57**, **3**, 255–289.

Valiela, I., Collins, G., Kremer, J., Lajtha, M., Geist, M., Seely, B., Brawley, T., and Sham, C. H. (1997). Nitrogen loading from coastal watersheds to receiving estuaries: Review of methods and calculation of loading to Waquoit Bay. *Ecology Applications*, **7**, 358–380.

Van der Laan, M. J., Pollard, K. S., and Bryan, J. (2003). A new partitioning around medoids algorithm. *J. Stat. Comput. Sim.*, **73**, **8**, 575–584.

Van der Maarel, E. (2007). Transformation of cover-abundance values for appropriate numerical treatment – alternatives to the proposals by Podani. *J. Veg. Sci.*, **18**, 767–770.

Van Wijk, R. (1989). Ecological studies on *Potamogeton pectinatus*. nutritional ecology, in vitro uptake of nutrients and growth limitation. *Aquat. Bot.*, **35**, 319–335.

Wagon, S. (2000). *Mathematica in Action, 2nd edition.* Springer Verlag, New York.